This is the author version published as:


Copyright 2010 Elsevier Ltd
Dynamic Visual Features for Audio-Visual Speaker Verification

David Dean, Sridha Sridharan

Abstract

The cascading appearance-based (CAB) feature extraction technique has established itself as the state of the art in extracting dynamic visual speech features for speech recognition. In this paper, we will focus on investigating the effectiveness of this technique for the related speaker verification application. By investigating the speaker verification ability of each stage of the cascade we will demonstrate that the same steps taken to reduce static speaker and environmental information for the visual speech recognition application also provide similar improvements for visual speaker recognition. A further study is conducted comparing synchronous HMM (SHMM) based fusion of CAB visual features and traditional perceptual linear predictive (PLP) acoustic features to show that higher complexity inherit in the SHMM approach does not appear to provide any improvement in the final audio-visual speaker verification system over simpler utterance level score fusion.

Key words: audio-visual speaker recognition; cascading appearance-based features; synchronous hidden Markov models

1 Introduction

Traditionally, the use of speech to recognise either words or speakers has been performed only in the acoustic modality. Whilst this area of research is fairly mature, there are still major problems with performance in real-world environments, particularly under high levels of acoustic noise. Audio-visual speech processing (AVSP) (covering both speech and speaker recognition) attempts to alleviate these problems through the addition of the visual modality to acoustic speech processing [1].

1 This research was supported by a grant from the Australian Research Council (ARC) Linkage Project LP0562101.
One of the most important factors in the final performance of an AVSP system is the choice of speech-based feature extraction techniques for the acoustic and visual modalities. While such feature-extraction research is very mature for the acoustic modality, the comparative novelty of visual speech processing has not yet resulted in a general consensus on the extraction of suitable visual features [2].

Because visual speech is fundamentally represented by the movement of the visual articulators, many feature extraction techniques focus on these movements rather than the stationary appearance within each frame. This approach has been shown to work very well for speech recognition [3], but it is not clear that it would apply for speaker recognition where static features, such as skin colour or facial hair, may be useful for identity purposes [4].

The extraction of visual speech features directly from the mouth region-of-interest (ROI) of a talking face has been shown to outperform geometric or contour-based feature extraction techniques for visual speech recognition [2]. However, a downside of these appearance-based feature extraction techniques is that each frame of the ROI contains a large amount of static speaker or environmental specific information that is unrelated to the movements of the visible articulators. Dynamic visual feature extraction techniques are designed to take the static ROI images and emphasize the dynamic nature of the visual speech over the stationary appearance within each frame. A number of techniques have been developed to extract these dynamic features, from simple approaches like difference images, or the use of delta and acceleration coefficients to more complicated techniques such as optical flow [1].

The current state-of-the-art in dynamic visual speech feature extraction is a multi-stage cascade of appearance-based (CAB) feature extraction techniques developed by Potamianos et al. [3], which has been shown to work well for speaker independent speech recognition [2]. While CAB features have been demonstrated for speaker recognition by Nefian et al [5], no detailed study on the effects of each stage of the cascade for speaker recognition has yet been conducted, and will therefore form the focus of this paper.

2 Dynamic visual speech features

While generally demonstrated to perform as well as or better than contour based methods [6,7], most simple appearance-based methods do tend to contain a significant amount of information irrelevant to the visual speech events. However, there are a number of techniques that have been demonstrated that attempt to extract the maximum visual speech information from the ROI, whilst discarding unwanted variance due to other factors. This section will be-
gin with a background of existing methods of dynamic visual speech features, and will then detail the visual feature extraction technique used for this paper.

2.1 Background

As visual speech is fundamentally represented by the movements of the visual articulators, the best features for representing visual speech are generally considered to focus on the movement of these features, rather than the stationary appearance within each frame [8,9]. While this is clearly the case in speech recognition applications, it is not completely clear that this would apply for speaker recognition, where the static features, such as skin colour or facial hair, of the ROI may be useful for identity purposes [4,1]. A number of researchers have shown that purely dynamic features can work well for speaker recognition applications [10,11,12], although there has not been any significant comparison of dynamic features with existing static features in the literature.

The simplest method of attempting to extract dynamic information from the video features is through the use of time-derivative-based delta and acceleration coefficients. These coefficients are generally used in addition to the original static feature values [13], although some researchers have discarded the static and used only the time-derivative features [14]. In a similar manner, rather than calculating frame differences using extracted features, the ROIs can be converted into frame-to-frame difference images before feature extraction can occur [15].

While time-derivative features, whether calculated before or after normal feature extraction, show the differences between adjacent frames, they do not directly indicate the movement of the visual articulators. For this purpose features based on calculating the optical flow [16] within the ROI have been used widely for both speech and speaker recognition applications in the visual domain [17,10]. However, it is not clear that there is any performance increase when compared to time-derivative-based features [15,17].

One technique that has recently shown good performance in AVSP applications is the use of LDA to extract the relevant dynamic speech features from the ROI through the CAB feature extraction process [18,2,5].

2.2 Cascading appearance-based features

An outline of the CAB feature extraction system used for this paper is shown in Figure 1, and can be seen to have three main stages:
Figure 1. Overview of the cascading appearance-based feature extraction system used for this paper. This is a simplified version of Potamianos et al.’s original cascade [3].

(1) Frames are (optionally) first normalised to remove irrelevant speaker or environmental information,

(2) Static features are extracted for each individual frame, and

(3) Dynamic features are generated from the static features over a window of several frames

In order to examine the utility of the CAB features for speaker verification, features will be extracted from both the static and dynamic stages of the cascade, with and without the frame normalisation. This will allow the usefulness of each stage of the cascade to be evaluated for the speaker verification task.

Frame normalisation

Before the static features can be extracted from each frame’s ROI, an image normalisation step is first performed to remove any irrelevant information, such as illumination or speaker variances. In Potamianos et al.’s original implementation of the cascade [3], this step was performed using feature normalisation after static feature extraction, but image normalisation has been shown to work slightly better due to the ability to handle variations in speaker appearance, illumination and pose as part of a wider pre-processing front-end [19]. As such, image mean normalisation was chosen over feature mean normalisation for these experiments.

This image normalisation step consists of calculating the mean ROI image over an entire utterance which can then be subtracted pixel-by-pixel from each ROI image before the ROI is presented to the static feature extraction stage.

The motivation behind normalising the ROI in this manner comes from the notion that a large amount of static speaker- and session-specific appearance-based information is collected in the standard appearance-based feature extraction techniques [20], and that this information would not be useful for modelling speech events. Of course, it is quite possible that this information would be useful for the speaker recognition application, so a version of the cascade will also be tested without this normalisation step to investigate this effect.
Figure 2. Most of the energy of a 2D DCT resides in the lower-order coefficients, and can be collected easily using a zig-zag pattern[21].

Static feature extraction

Once the ROI has been (optionally) normalised, static visual speech features can then be extracted. The main aim of feature extraction is to provide compression of the raw pixel values in the ROI whilst still maintaining good separation of the differing speech events. Discrete cosine transform (DCT) based feature extraction was chosen for Potamianos et al.’s original cascade [3] as well as this implementation in this paper, as they can more easily be calculated than other feature extraction techniques such as PCA [2].

For the static feature extraction stage of this cascade, the top $D^S$ coefficients were taken in a zig-zag pattern [21] in the two-dimensional DCT of the ROI, as shown in Figure 2. For the evaluation of the static features, delta and acceleration components were added to result in a $3 \times D^S$ dimensional feature space, but only the primary $D^S$ features were used as input to the dynamic feature extraction stage.

The frame-normalised version of these DCT features will be referred to as mean-removed DCT (MRDCT) features throughout the remainder of this paper.

Dynamic feature extraction

To extract the dynamic visual features that have been shown to improve human perception of speech, this stage of the cascade extracts linear discriminant analysis (LDA) based features over a range of consecutive ROIs. By operating over a number of consecutive frames centred on each frame under consideration, the LDA stage emphasizes the dynamic over the static features of the visual speech. The input to the LDA algorithm for the concatenated ROI features around $o^S_t$ is therefore given as

$$o^C_t = [o^S_{t-J}, \ldots, o^S_t, \ldots, o^S_{t+J}]$$ (1)

Where $o^S_t$ is the static video features at time $t$ and $J$ is the number of frames being concatenated on each side of the central frame. It can be seen that this results in a feature vector of size $D^C = (2J + 1)D^S$. 
Once the LDA transformation matrix was calculated using training data, it can then be used to transform the static observation vectors from the DCT stage of the cascade to form the dynamic visual speech features used to train and test the models for speaker verification. The final dynamic feature vector dimensionality can be reduced by only choosing the first $D^D$ eigenvectors from the calculated LDA transformation matrix before transforming the concatenated static features.

### 3 Synchronous HMMs

In terms of modelling the relationship between the audio and visual modalities, SHMMs can be seen as providing a middle ground for audio-visual speech processing between feature fusion and asynchronous HMMs [2]. Unlike feature fusion, SHMMs can model the reliability of each stream independently, but they cannot model the asynchronicity between the two streams as asynchronous HMMs can [23]. However, the small performance benefit of modelling the asynchronicity may not be worth the increase in model complexity, such as in embedded environments where processing power or memory may be limited.

A SHMM can be viewed as a regular single-stream HMM, but with two observation-emission Gaussian mixture models (GMMs) for each state—one for audio, and one for video—as shown in Figure 3(b). In the existing literature, SHMMs have been trained in one of two manners: Two single-stream
HMMs can be trained independently and combined, or the entire SHMM can be jointly-trained using both modalities. Because the combination method makes an incorrect assumption that the two HMMs were synchronous before combination, better performance can be obtained with the joint-training method [18]. Recently an additional method of training SHMMs, Fused-HMM (FHMM) adaptation, was introduced by the authors of this paper and has shown promise for audio-visual speech processing tasks [24].

3.1 Background

Given the audio and visual observation vectors \( o_{a,t} \) and \( o_{v,t} \), the observation-emission score of SHMM state \( u \) is given as

\[
P(o_{a,t}, o_{v,t} | u) = P(o_{a,t} | u)^{\alpha} P(o_{v,t} | u)^{1-\alpha}
\]

where \( \alpha \) is the audio stream weighting parameter \( 0 \leq \alpha \leq 1 \), with the corresponding video stream weighting parameter being \( 1 - \alpha \).

The SHMM parameters can then be defined as \( \text{\( \lambda_{av} = [\lambda_{av}, \alpha] \)} \) where \( \lambda_{av} = [A_{av}, B_{a}, B_{v}] \). In the underlying HMM parameters \( \lambda_{av} \), the joint state-transition probabilities are contained in \( A_{av} \), and \( B_{a} \) and \( B_{v} \) represent that observation-emission probability parameters of the audio and video modalities respectively [2]. Training of the SHMM is the process of estimating these parameters.

3.2 Joint training

The joint-training process estimates the parameters in \( \lambda_{av} \) using Baum-Welch re-estimation on both the audio and video streams simultaneously. The Baum-Welch re-estimation algorithm is the iterative process used to calculate the SHMM parameters from a training set of representative speech events, and briefly be outlined for SHMMs as follows:

1. Use the SHMM parameters (emission and state transition likelihoods) and the training data to estimate the state-occupation probability \( L_j(t) \) for all states \( j \) and times \( t \).
2. For each stream, use the state-occupation probability and the training data to re-estimate new SHMM parameters.
3. Repeat at Step 1 if the SHMM parameters have not converged.

As the Baum-Welch algorithm requires a initial set of SHMM parameters to form the first estimate of \( L_j(t) \), the parameters are generally initialised by
segmenting the training observations equally amongst the state models. From these segmented training observations the initial set of observation-emission parameters are determined for each state. From this point, the Baum-Welch algorithm can take over to refine the state-alignments and SHMM parameters until they have converged upon a solution.

However, while this approach can be easily implemented for speech recognition applications, the existing version (3.4) of the HMM Toolkit [25] does not have good support for speaker-adaptation of SHMMs, limiting the easy application of jointly-trained SHMMs to speaker recognition applications.

3.3 FHMM adaptation

FHMMs were introduced as an alternative to other multi-stream modelling techniques, designed to maximise the mutual information between the two modalities. As originally implemented FHMMs consisted of a continuous HMM for the dominant modality combined with a discrete vector-quantisation classifier for the subordinate modality within each state [22], as shown in Figure 3(a). The subordinate classifiers were trained based on the forced-alignment of the dominant HMM on the training set. This original design was extended by the present authors in [26] to improve the modelling of the subordinate modality by using a continuous classifier.

The FHMM-adaptation process results in two continuous GMMs inside each state of the original dominant HMM, which can be seen to be identical to the multi-stream model shown in Figure 3(b). Therefore it can be concluded that rather than being an alternative model type, FHMMs can be regarded as an alternative way of training a regular SHMM by adaptation from the dominant single-stream HMM rather than jointly-training on both modalities. The choice of the dominant modality for FHMM-adaptation should be based on the more reliable modality, which for speech processing will generally be the acoustic one [26].

FHMM adaptation can be considered as closer to the independent estimation method of SHMM parameters introduced briefly in Section 3 than joint training. However instead of estimating the audio and video HMM parameters independently and combining them to make the final SHMM, the FHMM-adaptation method uses the state alignments from a previously estimated audio HMM to directly train the video observation-emission likelihood parameters for each state, and combines them to make a SHMM. By using the audio alignments, it can be ensured that there is no problem with the states not being synchronous in training. Training using FHMM-adaptation is also a shorter process than both separate and joint-estimation of SHMMs, as Baum-
Table 1  
XM2VTS dataset configurations used in these experiments

Welch re-estimation only has to occur for the audio stream.

In our previous work we have shown that this FHMM-adaptation method can provide significant improvement over joint-training for the speech recognition application over a wide range of acoustic noise levels [24]. While we could not directly compare the speaker verification ability of jointly-trained and FHMM-adapted SHMMs due to limitations outlined previously, we believe that the FHMM-adapted SHMM models trained using this method should be comparable to jointly-trained SHMM models, if not better.

4 Experimental setup

4.1 Training and testing datasets

For this experiment, training, testing and evaluation data were extracted from the digit-video sections of the XM2VTS database [27]. The training and testing configurations used for these experiments was based on the second configuration of the XM2VTSDB protocol [28], but adapted to allow more tests than provided by the standard protocol. Each of the 295 speakers in the database has four separate sessions of video where the speaker speaks two sequences of two sentences of ten digits. In each of the configurations, two sessions were used for training, one for evaluation and one for testing, allowing for 12 configurations in total, as shown in Table 1. By comparison, the XM2VTSDB protocol only allows for two distinct configurations. While this approach may result in some temporal anomalies when compared to the XM2VTS protocol (such as a smaller average test-train temporal distance, and training on future data), these were viewed as unlikely to affect the experiments significantly, and worth the large increase in evaluation experiments.

These experiments were performed as verification experiments, where the speaker would attempt to enter the system by claiming the identity of a particular client. To perform this task, the speakers were split into two groups: clients, who claimed their own identity; and impostors, who claimed the identity of one of the clients.

As per the XM2VTSDB protocol, 200 speakers were designated clients, and 95 were used as impostors. For each client testing sequence (2 per session), 20
sequences were chosen at random from the impostor set allowing for a total of 400 (200 × 2) client tests and 8000 (200 × 2 × 20) impostor tests for each configuration. Over all 12 configurations, 4800 client tests and 96000 impostor tests are performed.

4.2 Feature extraction

In order to provide a baseline for the visual speaker verification experiments, perceptual linear prediction (PLP) based cepstral features were extracted from the acoustic speech. Each acoustic feature vector consisted of the first 13 PLPs including the zeroth, and the first and second time derivatives of those 13 features resulting in a 39 dimensional feature vector. These features were calculated every 10 milliseconds using 25 millisecond Hamming-windowed speech signals.

Visual features were extracted from a manually tracked lip region-of-interest (ROI) from 25 fps (40 milliseconds / frame) video data. Manual tracking of the locations of the eyes and lips were performed every 50 frames, and the remainder of the frames were interpolated from the manual tracking. The eye locations were used to normalise the in-plane rotation of the lips. A rectangular region-of-interest, 120 pixels wide and 80 pixels tall, centered around the lips was extracted from each frame in the video. Each ROI was then reduced to 20% of its original size (24 × 16 pixels) and converted to grayscale.

The choice of a fixed size ROI was originally made for our earlier speaker-independent speech recognition research, and was designed to simulate the effect of a bank of fixed size Viola-Jones[29] lip detectors. Manual inspection of a large segment of the database has shown this approach works well in extracting consistent lip images from the XM2VTS database, but extraction of lip ROIs from less consistent data sources will likely require an additional scale normalisation stage.

Following the ROI extraction, the mean ROI over the utterance is optionally removed as the first stage of the CAB feature extraction. The top 20 static DCT-based features are then extracted as the second stage, and deltas and acceleration coefficients were added, resulting in 60 dimensional video feature vectors. Subsequently, to incorporate dynamic speech information, 7 neighboring such features (without the temporal-derivative coefficients) over ±3 adjacent frames were concatenated, and were projected via LDA to 20 dimensional dynamic visual feature vectors. The delta and acceleration coefficients of this vector were then incorporated, resulting in 60 dimensional visual feature vectors at the final stage of the cascade.

As a result of the application of the CAB feature extraction process, 4 video
feature representations were available for testing the speaker verification system. Including the acoustic PLP representation results in 5 feature representations tested for speaker verification:

- acoustic PLP features (A-PLP)
- static video without normalisation (V-DCT)
- static video with normalisation (V-MRDCT)
- dynamic video without normalisation (V-LDA-DCT)
- dynamic video with normalisation (V-LDA-MRDCT)

The A-PLP, V-DCT and V-MRDCT features were designed such that they could be extracted from any given utterance without any prior knowledge of the type of data they were working with, allowing their feature vectors to be used for each of the configurations of the XM2VTS database. However, because the LDA-derived features, V-LDA-MRDCT and V-LDA-DCT were trained based on acoustic speech events in the training sessions of the framework, each unique training configuration of the framework had to use a differing set of LDA-derived visual feature vectors. As a result, each sequence being tested had 6 different feature representations for V-LDA-MRDCT and V-LDA-DCT based upon which XM2VTS configuration was being tested.

4.3 Speaker verification

For the experiments conducted in this paper, text-dependent speaker verification was performed by aligning the speaker and background word-based models according to the known transcriptions for the testing sessions on each of the 12 configurations of the XM2VTS database. Each speaker-dependent model was tested alongside the background speech model for both the correct and a selection of impostor speakers and the difference of the two scores were plotted on detection error trade-off (DET) plots [30]. As the fusion experiments conducted in the second half of this paper were performed over a number of noise levels, the results of these experiments will be reported as the equal error rate (EER) of each of these DET plots to allow the relative performances to be compared easily in the limited space available.
<table>
<thead>
<tr>
<th>Model</th>
<th>Mixtures</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-PLP</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>V-DCT</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>V-MRDCT</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>V-LDA-DCT</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>V-LDA-MRDCT</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2

Best performing HMM topologies chosen for each datatype for the visual speaker verification experiments.

5 Visual speaker verification

5.1 Speaker modelling

In order to test the ability of each stage of the visual feature extraction cascade, text-dependent speaker models were trained and tested against all 12 configurations of the XM2VTS database defined earlier. These speaker-dependent models were generated by adapting word-based background HMMs to each individual speaker. The original background HMMs were themselves generated using the HMM Toolkit [25] to train HMMs using the training sequences for each configuration of the XM2VTS database over both clients and impostors. These models were then adapted to each individual client speaker’s training sequences using maximum a posteriori (MAP) adaptation [31]. Empirical experiments were performed on a single configuration to determine the best topology for each datatype, which are shown in Table 2.

5.2 Results and discussion

The results of the visual speaker verification experiments are shown in Figure 4, and from the results shown there, it can be seen that both the frame normalisation (DCT vs MRDCT) and application of speech based LDA within the cascade provide a benefit to speaker verification. In both cases, the lowest error rates occur when both the frame normalisation and LDA stages of the cascade are applied, resulting in the V-LDA-MRDCT video features.

Interestingly, these experiments show that the same features that have been shown to perform well for the task of speech recognition by other researchers [2] also perform very well for speaker verification. All of the stages of the cascade...
that improved visual speech recognition by downplaying the static speaker and session specific information also provided similar benefits for the speaker verification experiments, even though the original intent of the cascade was to provide a form of normalisation across speakers and subsequently improve speech recognition in unknown speakers.

These results suggest that for visual speaker recognition, the dynamic behavioural nature of speech could be at least as important, and possibly more-so than the more static physiological characteristics [32]. That is, it may be easier to recognise speakers by how they speak, than by their appearance while they speak. This also has the benefit that as static appearance is less important, environmental conditions such as illumination, and within speaker variations such as facial hair or makeup, become less of an issue provided they do not change throughout an utterance, and as long as the extraction of dynamic features can still be performed adequately.

6 Audio-visual speaker verification

Having established the primacy of the CAB visual feature extraction methods presented in the previous section, the best performing V-LDA-MRDCT video features were used to train and test an audio-visual speaker verification system alongside the traditional PLP-based acoustic features. Additionally, in order to demonstrate the ability of dynamic video features to improve audio-visual speaker verification in poor acoustic conditions, background office-babble style noise was artificially added to the audio tracks at levels of 0, 6, 12 and 18 dB of signal-to-noise ratio (SNR).
<table>
<thead>
<tr>
<th>Model</th>
<th>Mixtures</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Fusion HMM</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>FHMM-Adapted SHMM</td>
<td>11</td>
<td>8 (audio), 16 (video)</td>
</tr>
</tbody>
</table>

Table 3

*Best performing HMM topologies chosen for the discriminative fusion HMM and the FHMM-adapted SHMM.*

6.1 **Baseline systems**

In order to adequately study the performance of the SHMM-based audio-visual speaker verification system, four baseline systems were also trained and tested for comparison:

- Audio-only (A-PLP) HMMs
- Video-only (V-LDA-MRDCT) HMMs
- Discriminative feature-fusion HMMs
- Output fusion of audio and video HMMs

The audio and video HMMs used for unimodal and output score fusion were taken directly from the earlier unimodal speaker verification experiments, and the discriminative feature fusion HMMs were trained in a similar manner to the unimodal HMMs. The topology of the feature-fusion HMMs was determined empirically on the first evaluation configuration of the XM2VTS database and is shown in Table 3.

For the training and testing of the discriminative feature-fusion system, the closest video feature vector was chosen for each audio feature vector and appended to create a single 99-dimensional static feature-fusion vector. No interpolated estimation of the video features between frames was performed. Five neighboring features over $\pm 2$ adjacent frames were concatenated, and were projected via an *inter*-frame linear discriminant analysis (LDA) cascade to 24 dimensional discriminative feature-fusion vector. The delta and acceleration coefficients of this vector were then incorporated, resulting in the final 72-dimensional feature-fusion vector. This process can be seen to be very similar to the extraction of dynamic video features outlined in Section 2 and was found to provide better feature fusion performance over simple concatenation.

In a similar manner to the unimodal HMMs, the speaker dependent feature-fusion models were generated by adapting separately trained background models to each of the individual speakers by MAP adaptation.

The output fusion baseline system was achieved through weighted-sum fusion of the unimodal acoustic and video HMMs scores over the entirety of each test utterance. Both the scores from the output fusion baseline and the FHMM-
adapted SHMMs were stream normalised to ensure optimal performance. More
detail on stream normalisation and weighting is provided in Section 6.3.

6.2 Fused HMM adaptation

Our FHMM method of adapting a SHMM from an audio-only HMM, adapted
from Pan et al.’s original implementation [22], can be viewed as a relatively
simple two step process:

1. For each audio training observation, we find the best hidden-state align-
ment of the audio HMM by force-aligning the training transcriptions.
2. We next train additional video GMMs for each state based on the video
observations lining up with the best hidden-state alignment in (1)

The FHMM-adapted SHMM used in these experiments was based on the base-
line acoustic HMM. Once the video observations that overlapped a particular
state in the acoustic HMM were determined, a 16 mixture GMM was trained
on those observations and the video GMM was added next to the state’s al-
ready existing acoustic GMM. Once this had been performed for each state
in each acoustic HMM, the result was a new set of SHMMs with the same
states as the audio HMMs, but containing GMMs for both acoustic and visual
information, as indicated in the final topology shown in Table 3.

In a similar manner to the baseline HMMs, the speaker dependent SHMMs
were speaker-adapted from background models to each of the individual speak-
ers. This speaker-adaptation process was performed similarly to the FHMM-
adaptation process outlined above, but instead of training the video GMMs di-
rectly on the video observations for each speaker, the speaker-adapted GMMs
were MAP adapted from the background video GMMs from the already
FHMM-adapted background SHMMs.

6.3 Stream weighting and normalisation

In order to maximise the performance of the resulting speaker verification
systems, a normalisation and weighting stage was performed on both the
FHMM-adapted SHMM and the output fusion design. This stage was not
possible in the discriminative fusion system due to the early combination
of the two streams. To perform the stream normalisation, a form of zero-
normalisation [33] was used for both the output fusion and SHMM systems.
For output fusion, this normalisation was performed over an entire utterance,
whereas the SHMM normalisation was performed on every frame, within the
SHMM states, during decoding. Further information on the SHMM normali-
sation process can be found in our earlier paper [34]. In order to allow every audio frame to have a corresponding video frame, the video frames were up-sampled to the same rate as the audio features with no calculated interpolation performed.

Once both the audio and video streams classifiers were normalised a number of experiments were performed using the evaluation sessions of the XM2VTS database to determine the best stream weights for the speaker verification experiments. In order to limit the search space, the stream weights were defined in terms of a single weighting parameter $\alpha$ representing the acoustic stream weight, with the video stream weight being defined as $1 - \alpha$. The results of these experiments are shown in Figure 5(a) for the output fusion system and in Figure 5(b) for the FHMM-adapted SHMM. Because output fusion only combines scores at the utterance level rather than the frame level of SHMM stream weighting, the output fusion tuning could easily be performed over all 12 configurations of the XM2VTS database while the SHMM tuning was limited to a single configuration due to computational constraints.

However, regardless of the comparative granularity of the tuning experiments shown in Figure 5, it can be seen that both the output fusion and SHMM systems responded similarly to the stream weighting parameter. Accordingly, a final weighting parameter of $\alpha = 0.2$ was chosen for the speaker verification experiments performed in this paper based on the average EER performance over all noise levels.

Figure 5. Performance of output fusion and FHMM-adapted SHMM systems as $\alpha$ is varied from 0 to 1.
6.4 Results and discussion

The results of the unsupervised FHMM-adapted speaker verification experiments are shown in comparison to the baseline systems in Figure 6. While it can be seen that the FHMM-adapted SHMMs perform well in comparison to early integration, it nonetheless performs catastrophically in noisy conditions, and is easily bested by the output score fusion of the uni-modal HMMs for all of the acoustic conditions under test. This poor performance of the feature fusion system in comparison to the output fusion approach is consistent with existing audio-visual speaker recognition research, and is believed to be related to the extra complexity and synchronicity issues involved in training feature-fusion speech models [1].

While it is certainly possible that an adaptive fusion approach could be used to allow the stream weighting parameters to be varied based on an estimation of the prevailing environmental conditions, a similar approach could also be applied for the output score fusion of the uni-modal HMMs.

Indeed, in order for SHMM approaches to speaker verification to improve over simple output score fusion of uni-modal HMMs, the SHMM approach would have to show that it can take advantage of some temporal dependency between the speech features. Because output-score fusion can only occur at the end of an utterance, or through an externally determined segmentation process, it would not be able to take advantage of the differences on a frame-by-frame basis.

However, at least for text-dependent speaker verification, such a situation is unlikely to occur, as the main role of the HMM structure is to align the GMMs within the HMM against the pertinent speech events. However, earlier experiments regarding the joint-training of SHMMs [34] have shown that both the acoustic and visual modalities are equally good at determining the hidden state boundaries of a known transcription. Therefore either acoustic or visual HMMs can align state-models equally well when evaluating a known phrase for text-dependent speaker verification, and limited benefit arises from aligning the states with both modalities.

7 Conclusion and future work

This paper investigated the CAB feature extraction process introduced by Potamianos et al. [3] to improve speech recognition for the related task of speaker verification. The experiments conducted within this paper found that this process also provided considerable improvement for speaker verification,
even though the aim of the CAB process was to remove or downplay the speaker (and environmental) specific information in order to improve the recognition of speaker-independent speech. That the speaker verification experiments improved in performance as static information was removed suggests that dynamic visual information can play a very important role in visual (and audio-visual) person recognition, particular when the facial movements are speech related. Of course, face recognition is a very mature area of research that has shown that static recognition of faces can provide good performance, and the possibility certainly exists of using a combination of static face and dynamic features to represent the visual modality with a minimum loss of information. Some promising versions of such systems have been developed [5], but this area is still a relatively new area of research.

While the more dynamic features extracted through the CAB feature extraction process provided better performance than static visual feature extraction techniques, fusion experiments with FHMM-adapted SHMMs found that no benefit was obtained by combining the dynamic video and acoustic features at the frame level. Rather, best performance was obtained by combining two independent unimodal classifiers at the whole utterance level through normalised and weighted output score fusion. This approach also had the significant advantage of simplicity, as the stream weights can be easily changed separate from the HMM decoding process, rather than within it as would be required for the SHMM-based approach.

While the XM2VTS database is one of the largest audio-visual speech databases that is available for to the general research community it is still a relatively small database compared to the many large speech databases available to the audio speech recognition community [35,36]. The XM2VTS database
is also limited in that all speech is collected in clean conditions (both acoustic and visual), which does not easily allow the utility of audio-visual speech processing to be evaluated in the conditions where it may be of most use compared to solely acoustic approaches. A number of audio-visual databases have been relatively recently released that may be useful for evaluating the work presented in this paper over adversarial conditions, such as the BANCA [37] database covering office conditions, and the AVICAR [38] in-car audio-visual speech database.

A possible avenue of future research that may be able to take advantage of the SHMM structure for speaker verification could focus on using the SHMM for limited-vocabulary text-independent speaker verification. By taking advantage of the SHMM’s ability to find the correct transcription through the network, as exhibited in our earlier work on speech recognition experiments [24], the SHMM may be able to perform well for speaker verification when the speech itself is unknown. While this could provide better performance than two unimodal HMMs in a similar configuration, it is not clear if this approach would be better than considering the output of two large-vocabulary text-independent speaker verification models in each modality, for which preliminary research has already shown CAB video features should perform well [39].

8 Acknowledgments

The research on which this paper is based acknowledges the use of the Extended Multimodal Face Database and associated documentation. Further details of this software can be found in [27] or at http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/.

References


techniques,” in Computer Vision and Pattern Recognition, 1992. Proceedings

audio-visual data,” in Joint AMI/PASCAL/IM2/M4 workshop, Martigny, CH,
2004.


University of Technology, Brisbane, Australia, 2007.

between different feature extraction techniques for audio-visual speech


[22] H. Pan, S. Levinson, T. Huang, and Z.-P. Liang, “A fused hidden markov model
with application to bimodal speech processing,” IEEE Transactions on Signal


[24] D. Dean, P. Lucey, S. Sridharan, and T. Wark, “Fused HMM-adaptation of multi-
stream HMMs for audio-visual speech recognition,” in Interspeech, Antwerp,

V. Valtchev, and P. Woodland, The HTK Book, 3rd ed. Cambridge, UK:
Cambridge University Engineering Department., 2002.

continuous fused HMMs,” in HCSNet Workshop on the Use of Vision in HCI
(VisHCI), 2006.

The extended M2VTS database,” in Audio and Video-based Biometric Person
Authentication (AVBPA ’99), Second International Conference on, Washington


[29] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of
simple features,” in Computer Vision and Pattern Recognition, 2001. CVPR


